Preparing for GWAS: Working with Data

Day 2 Team

Alesha Hatton, Anaïs Thijssen, Brad Verhulst, Daniel Hwang, Dinka Smajlagic, Elizabeth Prom-Wormley, Geng Wang, Gunn-Helen Moen, Jose Morosoli, Kristen Kelly, Madhurbain Singh, Penelope Lind, Penghao Xia, Sophie Breunig, Tunde Olasege

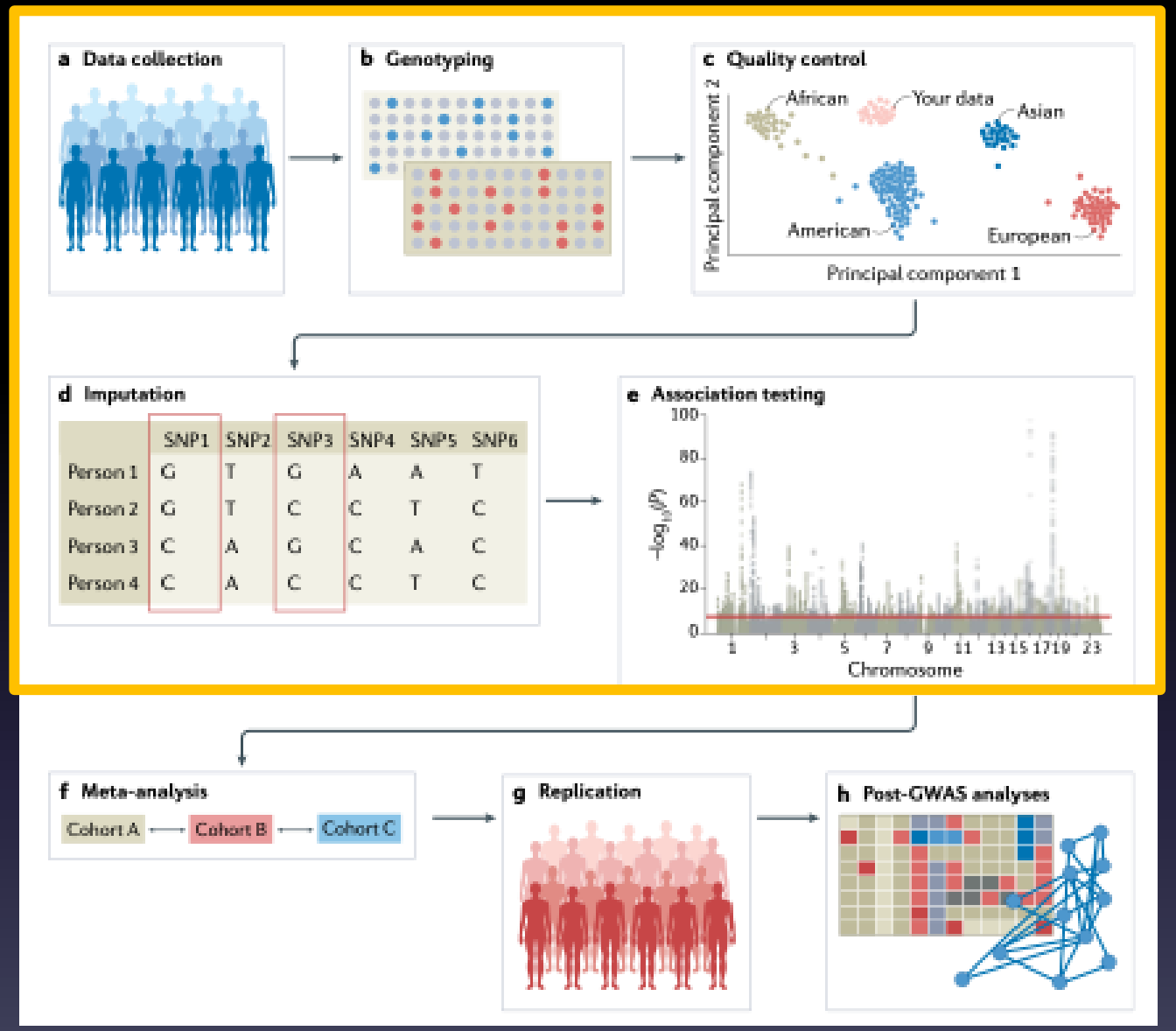
With Special Thanks to Sarah Medland, Katrina Grasby, and Lucia Colodro Conde

June 2, 2026

https://www.colorado.edu/ibg/workshop-2026/syllabus/working-data

Day 2 Objectives

- By the end of today's session you should be able to:
  - Prepare phenotypic and genotypic data to conduct a basic GWAS with few-/no- assumption violations
  - Assess the degree to which phenotypic and genotypic data are ready for use in GWAS using PLINK and R
  - Summarize the phenotypic and genotypic data-related assumptions of a GWAS and their implications when violated
  - Identify strategies to address GWAS assumption violations



1) Basic GWAS data and assumptions (single-sample study)

2) Data considerations where assumptions may not be fully achieved (single- sample study)

3) Data preparation considerations for a multi-sample meta-analysis

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews. Methods Primers*, *1*, 1-21.Article 59. https://doi.org/10.1038/s43586-021-00056-9

Overview of Basic PLINK Files

*.ped
Individuals and their genotypes

*.fam
Information on the individuals

Covariate file is separate

*.map
Location details on the genetic markers based on the current genome build

*.bim
Genetic marker info-Possible alleles

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):e1608. doi:10.1002/mpr.1608

Phenotypic Data Prep

- Inclusion / exclusion criteria
  - Data collection details
- File Format
  - File type (PLINK or other)
  - Header, delimiter, missing-value codes

Basic GWAS Phenotypic Assumptions

- ***Continuous data have a normal distribution:*** Outcome should have a normal distribution if you are conducting linear regression

- ***Binary Data have an equal case-control ratio***: The number of cases is close to the number of controls (~1 case : 1 control)

Basic GWAS Genotypic Data Assumptions

- ***No SNP missingness***: At the participant level, there are not a large proportion of SNPs that are missing
- ***No Sex Discrepancies***: Self-reported sex and sex from X chromosome heterozygosity are the same
- ***Appropriate Minor Allele Frequency Threshold***: SNPs above a set MAF threshold
- ***SNPs in HWE***: Included SNPs are in Hardy Weinberg Equilibrium
- ***Participants with Some Heterozygosity***: At the participant level, the proportion of SNPs with Low-/High- rates of Heterozygosity is low.
- ***Non-Related Individuals***: Participants in the GWAS are not related to each other
- ***No Population Stratification***: Participants in the GWAS do not represent multiple subgroups with different allele distributions (e.g., genetic ancestry groups)

Most GWAS Do Not Uphold All Assumptions

Be Aware, Be Careful, Be Diligent

Document always and often

One Basic Assumption – No Covariates or Confounders

- ***No third variable influence***. The distribution of the outcome (or genotyping data quality) is not due to a third variable

- Classic Covariates- age, age^2, biological sex
- Laboratory/Study Design-Related Concerns
  - Processing/isolation of DNA
  - Genotyping batch effects
  - Project study site
  - Multiple genetic ancestry groups

Additional GWAS Data Files

Feature	PLINK 1 (1.07)	PLINK 1.9	PLINK 2
Status	Legacy / Retired	Fully stable and complete	Advanced alpha (active development)
Speed	Slow (often limited by memory)	1-4 orders of magnitude faster than PLINK 1	Even faster, highly parallelized multithreading
Native File Formats	`.bed`, `.bim`, `.fam`	`.bed`, `.bim`, `.fam`	`.pgen`, `.pvar`, `.psam` (faster I/O)
Allele Definitions	Biallelic (A1/A2, determined by frequency)	Biallelic (A1/A2, determined by frequency)	Multiallelic natively supported (REF/ALT)
Missing Data / Dosage	Lossy import of dosage/probabilities	Lossy import of dosage/probabilities	Retains genotype likelihoods & phased states

Other file formats

Variant Call Formats (`.vcf`, `.bcf`) for sequencing data

Oxford format (`.gen`, `.bgen`, `.sample`)

QC Step	Check	Rationale	Analysis
Phenotype set-up	Verify phenotype coding (from analysis plan*)	Failure to address or align with other studies (when conducting meta-analysis) may result in incorrect interpretations	Check units for continuous traits
			Check case-control definitions and coding (from analysis plan*)
			Inclusion / exclusion criteria
			Covariates (from analysis plan*) Additional cohort-specific covariates (e.g., batch, scanner)
	File format		Header, delimiter, missing-value codes
	Duplicates / Repeated measures		Apply a rule for keeping samples (e.g., latest visit, case status)
	Related samples		Use appropriate GWAS analysis program
Binary Traits	Unbalanced case/control ratio	Extremely unbalanced ratios may increase Type I error (false positive)	Use appropriate GWAS analysis program
Continuous Traits	Distribution	Failure to assess and address or align with other studies (when conducting meta-analysis) may reduce the sensitivity of the measurement and subsequent power to detect real effect of genetic association	Use histogram/box plots to check for outliers and skewness
			Check for normality
	Transformation		Transform if not normal.
			Ensure all cohorts use the same transformation to keep the same scale.*
	Outliers		Remove implausible values
			Remove if >3 SD / winsorize
	Binning		Convert to bins / convert bins

Phenotypic (and Covariate) Data Checklist to Test/Address Assumptions

*Note: Pay attention to analysis plan/consortium guidance if conducting meta-analysis

Genotypic Data Checklist to Test/Address Assumptions

QC Step	Check	Rationale	Analysis
Genotype set-up	Confirm genome build	Ensure proper SNP location alignment	Liftover (if needed)
	Harmonize across batches		Update CHR, BP & SNP names
	Genotype File Format	Match file to program	VCF, BGEN, .bed, .bim, .fam
SNP-level QC	Overall missingness	Ensure small proportion of SNPs that are missing	Remove if > 5%
	Minor allele frequency	Most GWAS are underpowered to detect associations with SNPs with a low MAF	Remove if < 0.5%
	Hardy Weinberg Equilibrium	Ensure use of SNPs with genotype and the allele frequencies that are constant over generations	Remove if P < 1e-6 , theshold may differ across samples in a meta-analysis
	Ambiguous SNPs	SNPs with indeterminate calls may result in incorrect interpretations	Remove strand-ambiguous SNPs

QC Step	Check	Rationale	Analysis
Sample-level QC	Overall missingness	High levels of SNP genotype missingness can indicate poor DNA quality or technical problems	Remove if > 5%
	Heterozygosity	High individual-level SNP heterozygosity might be due to low sample quality. Low levels of heterozygosity may result from inbreeding.	Remove if >3 SD
	Sex check	A discrepancy may reflect sample mix-ups in the lab	Update sex (if needed)
			Drop sample if sex error
	Relatedness	Without appropriate correction, the inclusion of data from relatives could bias estimations of SNP effect size standard errors	Remove if pihat ≥ .2 (if needed) or >0.05 if based on genetic relationship matrix
	Ancestry	Allele frequencies can differ between genetic ancestry subpopulations. Without appropriate correction, population stratification can lead to false positive associations and/or mask true associations.	Multidimensional Scaling (MDS)

Genotypic Data Checklist to Test/Address Assumptions

Breakout Room Recommendations

- Start with the phenotypic QC tutorial/RCR question (~20 min, 5 min break)
  - https://qimr.az1.qualtrics.com/jfe/form/SV_5AuuMePy7gDTeJg
- Once you arrive, decide who will:
  - Run code/share screen
  - Read out tutorial text (in case the coder cannot see)
  - Write answers in tutorial page
  - Keep time
  - All should provide feedback on answers for the tutorial page
- Bonus phenotypic tutorial at the end. Complete genotypic QC tutorial first.
- Proceed to genotypic QC tutorial (~1 hour, 20 minutes)
  - Link in the phenotypic QC tutorial https://qimr.az1.qualtrics.com/jfe/form/SV_3eXrS1WZX2gS39s

Questions?

- Talk to a tutor when they stop by

- Share your question in the workshop forum under Day 2

https://isgw-forum.colorado.edu/c/isg-workshops-2026/day-2/31

Take Home Points

- There are several steps in conducting QC of data prior to running a GWAS. These are crucial to ensuring your GWAS has the greatest chance of success and that results will be interpretable

- Phenotypic QC- Phenotypic and covariate data were generally concluded to be ready to be included into GWAS. If you want to see how we went from the raw version of these data to the version you worked with, try out the extended phenotypic QC tutorial

- Genotypic QC- You were able to